Eur. J. Clin. Chem. Clin. Biochem.
Vol. 30, 1992, pp. 405–414

Robust Bivariate Errors-in-Variables Regression and Outlier Detection

By *U. Feldmann*

Abteilung für Medizinische Statistik, Biomathematik und Informationsverarbeitung
Universität Heidelberg, Klinikum Mannheim, Germany

(Received January 13/April 28, 1992)

Dedicated to Professor Berthold Schneider in honour of his 60th birthday

Summary: A bivariate regression model is introduced where both variables are subject to error. The structural regression line is equivariant against interchanging coordinates and permits bivariate calibration, i.e. the prediction of one variable by means of the other. Maximum likelihood and robust parameter are estimated, based on order statistics. Residual analysis and outlier detection are performed. The model is applied to the comparison of clinical chemical analytical methods.

Introduction

Many sciences widely use calibration techniques in order to compare different measuring methods. For instance in clinical chemistry, the comparison of different analytical methods, measuring the same substance, the so-called analyte, is a problem of great concern. Formally, calibration consists in predicating a value y of a measurement Y, given the value x of the measurement X. Calibration is usually performed with the aid of regression analysis. However, ordinary regression analysis distinguishes between explanatory (or independent) and response (or dependent) variables. The former variables are assumed to be free of measurement errors, while only the response variables are assumed to be affected by such errors.

In many practical applications this assumption is apparently unjustified, thereby leading to biased results when ordinary regression models are applied. Another restriction of ordinary regression is that distributional assumptions must be made with respect to the response variable. Usually the response variable is defined to be normally distributed. This assumption may also be inappropriate in many practical applications.

In order to avoid the second restriction, so-called robust regression approaches are applicable. Suitable textbooks on robust statistics (1) and in particular on robust regression (2) are available. The aim of robust regression is to detect outliers, which very frequently occur in real data, and to adjust their influence in the fit of the data. There are two general approaches to robust regression. Firstly, methods which protect against outliers in y, such as *Huber*'s M-estimator (1), and secondly high-breakdown methods which protect against outliers in x and y, e.g. *Rousseeuw*'s least median of squares estimator (2). Applications of the latter method to analytical chemistry can be found in l.c. (3). These regression models, however, remain univariate as long as only the response variable is assumed to be subjected to measurement errors.

Bivariate errors-in-variables analysis is usually conducted with the aid of the structural relationship model, introduced by *Wald* (4). A comprehensive discussion of structural analysis may be found in l.c. (5). Applications to clinical chemistry are given, for example, in l.c. (6). The robust estimation of the structural line may be conducted with the aid of order statistics (7), partitioning of the data (8), and jackknife techniques (9). However, it should be emphasized that the structural line must not be confused with a bivariate regression line or a calibration line.

The bivariate high-breakdown robust regression model should have the following properties:

(i) both measurements are errors-in-variables,

(ii) the regression line is scale equivariant with respect to magnifications of the axes,

(iii) the regression line is equivariant against the interchange of axes,

(iv) bivariate calibration is permitted, i.e. the prediction of Y given a x-value, and simultaneously, the prediction of X given a y-value, and finally,

(v) a bivariate residual analysis and outlier detection are possible.

An approach, satisfying these five conditions will be termed as a structural regression model.

It should be noted that the conditions (iv) and (v) do not hold in the framework of structural relationship models. The basic concept of structural regression was developed geometrically by *Feldmann & Schneider* (10).

The biometrical definition of structural regression is outlined in Section 1 and the decomposition of bivariate errors-in-variables into a residual variable and a location variable is conducted in Section 2. Distributional and robust parameter estimation is investigated in Sections 3 and 4. In Section 5, residual analysis and outlier detection is considered, and the model is applied to the comparison of clinical chemical analytical methods in Section 6.

1. Bivariate Calibration

Assume (X, Y) to be bivariate random variables whose realizations (x, y) are observable measurements, e. g. of an analyte measured with two different methods. To define a bivariate linear regression model, predictor variables X′ and Y′ are introduced which are related to the observable variables X and Y by the linear equations:

$$Y' = \alpha + \beta X \text{ and } Y = \alpha + \beta X' \qquad \text{(Eq. 1)}$$

with a probability of one.

The straight line with intercept α and slope β is the bivariate calibration line, i.e. one can predict a realization of Y as $y' = \alpha + \beta x$, given the x-value, and simultaneously, predict the realization of X as $x' = (y - \alpha)/\beta$, given the y-value, by using the same regression line.

In order to identify the calibration line it is assumed that the predictor variables (X′, Y′) and the observable variables (X, Y) are equivalent up to the second order moments.

First we assume equivalent expectations, i. e.

$$\mu_{x'} = \mu_x \quad \text{and} \quad \mu_{y'} = \mu_y \qquad \text{(Eq. 2a)}$$

Taking Eq. 1 into account, this relates the intercept α to the first order moments of the observable variable

$$\alpha = \mu_y - \beta\mu_x, \qquad \text{(Eq. 2b)}$$

and the following relationships between the second order moments hold:

$$\begin{aligned} \sigma^2_{x'} &= \beta^{-2}\,\sigma^2_y \\ \sigma^2_{y'} &= \beta^2\,\sigma^2_x \\ \sigma_{x'y'} &= \sigma_{xy} \end{aligned} \qquad \text{(Eq. 2c)}$$

From Eq. 2c we also obtain

$$\sigma_{x'}\,\sigma_{y'} = \sigma_x\,\sigma_y. \qquad \text{(Eq. 2d)}$$

Hence, independent of any particular choice of the slope β, the covariances, the products of the standard deviations, and consequently the coefficients of correlation, $\varrho_{x'y'} = \varrho_{xy}$ where $\varrho_{xy} = \sigma_{xy}(\sigma_x\,\sigma_y)^{-1}$, are equivalent under assumption (Eq. 2a).

In order to identify the slope, additionally the variances are assumed to be equivalent:

$$\sigma^2_{x'} = \sigma^2_x \quad \text{and} \quad \sigma^2_{y'} = \sigma^2_y. \qquad \text{(Eq. 3a)}$$

Considering Eq. 2a, the assumption of Eq. 3a relates uniquely the slope β of the bivariate regression line to the second moments of the observable variable (X, Y):

$$\sigma^2_y = \beta^2\,\sigma^2_x \quad \text{or} \quad \beta = \text{sign}(\sigma_{xy})\,\frac{\sigma_y}{\sigma_x}. \qquad \text{(Eq. 3b)}$$

This slope β is well known in linear regression analysis. For example, it corresponds to the "SD-line" (l. c. (11), page 122) and is nothing else but the geometric mean of the slopes of the two ordinary regression lines from y to x and from x to y, respectively, seen from the X-axis. As early as 1970, *Averdunk & Borner* (12) proposed this line for the comparison of analytical methods in clinical chemistry. In the framework of linear structural relationships β can be obtained (6) as the slope of the standardized principal component.

However, although a special solution of the linear structural relationship is recognized, it should be emphasized that within the framework of structural regression this solution has a quite different interpretation, since is it obtained as the slope of a bivariate calibration line.

The connection to the linear structural relationship model is easily seen by summation and subtraction of equations 1:

$$Y^* = \alpha + \beta X^* \qquad \text{(Eq. 4a)}$$

where

$$Y^* = \frac{Y + Y'}{2} \text{ and } X^* = \frac{X + X'}{2}$$

$$E_y = -\beta E_x \qquad \text{(Eq. 4b)}$$

where

$$E_y = \frac{Y - Y'}{2} \text{ and } E_x = \frac{X - X'}{2}$$

and we have

$$X = X^* + E_x \text{ and } Y = Y^* + E_y \qquad \text{(Eq. 4c)}$$

In the framework of structural relationships (6), X* and Y* are considered as latent variables, indicating the hidden error-free 'true' measurements (x*, y*), and the equations 4a and 4c define the structural line. This line, however, is only identifiable if additional assumptions are made with respect to the error terms E_x and E_y. In the structural relationship model these errors are assumed to be stochastically independent of each other and of the respective latent variables. The standardized principal component is achieved if the ratios of the standard deviations of the error terms and of the observable variables are assumed to be identical, i.e. $\sigma_{Ey}/\sigma_{Ex} = \sigma_y/\sigma_x$.

In the structural regression approach, presented here, the errors are related by Eq. 4b and are dependent with probability 1. However, in contrast to the structural relationship approach, the error terms are not used to identify the slope of the structural regression line. The identification is achieved by condition Eq. 3a concerning the variances of the predictor variables.

2. Orthogonal Decomposition

To be able to conduct residual analysis, an orthogonal decomposition of the bivariate distribution (X, Y) into a residual variable U and a location variable V is performed:

$$U = \frac{\sqrt{1 + \beta^{-2}}}{2}\{Y - \mu_y - \beta(X - \mu_x)\}$$

and (Eq. 5)

$$V = \frac{\sqrt{1 + \beta^{-2}}}{2}\{Y - \mu_y + \beta(X - \mu_x)\}$$

These variables have zero expectations, i.e. $\mu_u = 0$ and $\mu_v = 0$, and if Eq. 3b holds, U and V are uncorrelated, i.e. $\sigma_{uv} = 0$. The assumptions of equality of the variances of the observable and predictor variables Eq. 3a, and of orthogonality of U and V, are equivalent.

Under condition Eq. 3b the variances of U and V become:

$$\sigma_u^2 = \frac{(\sigma_x^2 + \sigma_y^2)}{2}(1 - |\varrho_{xy}|)$$

and (Eq. 6)

$$\sigma_v^2 = \frac{(\sigma_x^2 + \sigma_y^2)}{2}(1 + |\varrho_{xy}|)$$

U is the residual variable of our model; it determines the position of a measuring point (x, y) with respect to the calibration line. For instance the realization $u = 0$ of U indicates a measuring point (x, y) which is located on the bivariate calibration line, i.e. $y = \alpha + \beta x$. Furthermore, a realization $u > 0$ indicates the point (x, y) located above the calibration line, i.e. $y > \alpha + \beta x$, while $u < 0$ indicates a point (x, y) below the calibration line, i.e. $y < \alpha + \beta x$.

|u| determines the distance between the measuring point (x, y) and the latent point (x*, y*). On the other hand |u| is the distance between the latent point (x*, y*) and the predicted point (x, y′) and (x′, y), respectively, whereas all three are located on the regression line according to Eq. 1 and Eq. 4a.

V is the location variable of our model determining the position of a latent point (x*, y*) on the calibration line. For instance, the realization $v = 0$ of V indicates that the latent point and the focal point coincide, i.e. $(x^*, y^*) = (\mu_x, \mu_y)$. Furthermore, $v > 0$ indicates a latent point located on the calibration line above the focal point, i.e. $x^* > \mu_x$ and $y^* > \mu_y$, while $v < 0$ describes a latent point beyond the focal point, i.e. $x^* < \mu_x$ and $y^* < \mu_y$. |v| is the distance between the focal point (μ_x, μ_y) and the latent point (x*, y*), both located on the calibration line.

Another relationship also holds:

$$\sqrt{1 + \beta^2}\,\{X - \mu_x\} = V - U$$

and

$$\sqrt{1 + \beta^{-2}}\,\{Y - \mu_y\} = V + U.$$

Interchanging the coordinates (X, Y) into (Y, X) replaces β by β^{-1} and U by $-U$, while V remains unchanged. Hence, the interchange of coordinates only produces a change of the sign of the residual variable. The geometrical interpretation is that $|v - u|$ is the distance between the predicted point (x, y') and the focal point (μ_x, μ_y), while $|v + u|$ is the distance between the predicted point (x', y) and the focal point.

3. Maximum Likelihood Estimation

Assume that sample points (x_i, y_i) for $i = 1, \ldots, n$ are independently drawn from a bivariate normal distribution (X, Y). Then the predicted points (x'_i, y'_i) are from the same distribution according to 1, 2a and 3a. The bivariate normal density distribution function $f(x'_i, y'_i)$ is used for maximum likelihood estimation, and it is shown in the Appendix that

$$\psi^2(\mu_x; \mu_y; \beta) = \sum_{i=1}^{n} |\beta|^{-1}(y_i - \mu_y)^2 + |\beta|(x_i - \mu_x)^2 \quad \text{(Eq. 7)}$$

is to be minimized with respect to the model parameters in order to obtain maximum likelihood estimates:

$$\psi^2(m_x; m_y; b) = \underset{\mu_x\ \mu_y\ \beta}{\text{Minimum}}\ \psi^2(\mu_x; \mu_y; \beta)$$

As derived in the Appendix the maximum likelihood estimates are:

$$b = \text{sign}(r_{xy})\frac{s_y}{s_x}$$

$$\text{and } m_x = \bar{x},\ m_y = \bar{y},\ a = \bar{y} - b\bar{x} \quad \text{(Eq. 8a)}$$

with standard errors of the slope estimate, b, and the intercept estimate, a:

$$s_b = |b|\sqrt{\frac{1 - r_{xy}^2}{n}}$$

and

$$s_a = s_y\sqrt{\frac{2(1 - |r_{xy}|)}{n}} \quad \text{(Eq. 8b)}$$

We will term this the structural regression (SR). The structural regression and the standardized principal component analysis lead to the same result for the regression and structural line (see Section 1). It should be emphasized, however, that only the concept of structural regression allows for the evaluation of unbiased standard error estimates (Eq. 8b) with respect to the slope and the intercept, while this is not valid (6) for the standardized principal component.

As a measure of closeness of the data points to the regression line a bivariate coefficient of determination can be expressed by

$$R^2 = \frac{2|r_{xy}|}{1 + |r_{xy}|} \quad \text{(Eq. 9)}$$

$R^2 = 1$ holds if, and only if, $|r_{xy}| = 1$ and $R^2 = 0$ is valid if, and only if, $r_{xy} = 0$. From equations 6 one gets $R^2 = 1 - s_u^2/s_v^2$. This definition is in accordance with the principles used in ordinary regression and has already been proposed in l. c. (10).

4. Robust Estimation

The aim is to robustify the parameter estimates of Section 3 by making use of order statistics. Instead of minimizing the sum of squares (Eq. 7), the median of squares can be considered

$$\psi_r^2(\mu_x; \mu_y; \beta) = \underset{i}{\text{med}}\,[|\beta|^{-1}(y_i - \mu_y)^2 + |\beta|(x_i - \mu_x)^2] \quad \text{(Eq. 10)}$$

and minimized with respect to the model parameters. This is a bivariate version of the so-called least median of squares estimator (LMS),

$$\varphi_r^2(\alpha; \beta) = \underset{i}{\text{med}}\,[(y_i - \alpha - \beta x_i)^2],$$

introduced by *Rousseeuw* (13) for ordinary robust regression. The calculation of the least median of squares estimates leads to a complicated discrete minimisation problem, whose numerical aspects were investigated in l. c. (14), and more recently in l. c. (15).

We will term Eq. 10 as a least median structural regression (LSR). The calculation of the parameters is more complicated than in the least median of squares regression since the least median structural regression determined three model parameters; a robust slope β as well as a robust focal point (μ_x, μ_y). In this paper the simplex algorithm (16) is used for function minimisation.

Tools to investigate statistical properties of the estimates, for instance to determine confidence intervals, are not available in the framework of least median of squares regression. The so-called reweighted least squares regression (l. c. (2), page 131) can be applied for the determination of statistical inferences. An analogous procedure is also applicable in the least median structural regression and will be derived in the next Section.

As an approximate estimator which avoids numerical difficulties and statistical shortcomings, a robust analogue to the structural estimator (Eq. 8a) is proposed, and this is called the absolute median structural regression (ASR). In this approach the ratio of the standard deviations in equation 8a is estimated by the median of the absolute ratio, and the slope and intercept estimates become:

$$b = \pm \underset{i}{\text{med}} \left\{ \frac{|y_i - \underset{k}{\text{med}}\,(y_k)|}{|x_i - \underset{k}{\text{med}}\,(x_k)|} \right\}$$

and (Eq. 11)

$$a = \underset{i}{\text{med}}\,(y_i - bx_i).$$

The sign of b corresponds to the sign of

$$\tilde{b} = \underset{i}{\text{med}} \left\{ \frac{y_i - \underset{k}{\text{med}}\,(y_k)}{x_i - \underset{k}{\text{med}}\,(x_k)} \right\}$$

The history of the estimator $\tilde{b}$ can be found in l. c. (2), pages 73 – 74. This estimator was investigated for the first time by *Hampel* (17), who then dismissed it, by stating that it may lead to a poor fit.

The absolute median structural regression yields an appropriate robust fit (see Application) and does not need any minimisation procedure, as the least median structural regression of equation 10 does. It is therefore easily conducted by standard software; only a program computing a median is necessary. It has the further advantage of permitting the determination of confidence intervals. Exact confidence limits of a median (l. c. (8), page 362) can be derived by applying the cumulative binomial distribution. We use the well known normal approximation, which is valid for $n > 30$.

Consider the ratios

$$b_i = \frac{|y_i - \underset{k}{\text{med}}\,(y_k)|}{|x_i - \underset{k}{\text{med}}\,(x_k)|} \qquad i = 1, \ldots, n$$

and assume that the values $b_{[j]}$ are ordered according to the index j (j = 1, ..., n). Then the [n/2] order statistic is the median and the slope estimate (Eq. 11) reads $b = b_{[n/2]}$, while the order $m = [n/2 - 1.96\sqrt{n/4}]$ indicates the lower bound $b_{[m]}$ of the 95% confidence interval of the slope β, and $b_{[n-m+1]}$ is the upper bound.

The robust 95% confidence interval $\{a_{[m]}, a_{[n-m+1]}\}$ of the intercept α can be evaluated analogously, considering the ordered values $a_{[j]}$ of $a_i = y_i - bx_i$. The intercept in Eq. 11 is estimated by $a = a_{[n/2]}$. The brackets [ ] denote the integer value of the argument.

5. Residual Analysis and Outlier Detection

According to equation 5, the residuals u_i and location points v_i are given by

$$u_i = \frac{\sqrt{1 + b^{-2}}}{2} \{y_i - m_y - b(x_i - m_x)\}$$

and

$$v_i = \frac{\sqrt{1 + b^{-2}}}{2} \{y_i - m_y + b(x_i - m_x)\}$$

They depend on the slope estimate b and the focal point estimate (m_x, m_y). In the least median structural regression the intercept is estimated by $a = m_y - b\,m_x$, and in the absolute median structural regression the robust focal point estimate is caluclated by

$$m_{rx} = \frac{A - a}{2b} \quad \text{and} \quad m_{ry} = \frac{A + a}{2},$$

with

$$a = \underset{i}{\text{med}}\,(y_i - bx_i)$$

and

$$A = \underset{i}{\text{med}}\,(y_i + bx_i)$$

In order to conduct outlier detection with respect to the residuals, robust scale estimation is performed according to l. c. (2), page 202. For that purpose a preliminary scale estimate s_u° is calculated:

$$s_u^\circ = 1.4826 \left(1 + \frac{5}{n - 1}\right) \sqrt{\underset{i}{\text{med}}\, u_i^2}$$

With this scale preliminary standardized residuals u_i/s_u° are calculated and used to determine weights w_i

$$w_i = \begin{cases} 1 \text{ if } \left|\frac{u_i}{s_u^\circ}\right| \leq 2.5 \\ 0 \text{ otherwise} \end{cases} \qquad i = 1, \ldots, n$$

The resulting robust variance estimate for the residual variable U is then calculated by the weighted sum of squares:

$$s_u^2 = \frac{\sum_{i=1}^{n} w_i u_i^2}{\sum_{i=1}^{n} \{w_i\} - 1}$$

The robust variance estimate is used for the detection of outliers. A certain point (x_k, y_k) may be flagged as an outlier if the corresponding standardized absolute residual $|u_k/s_u| > \lambda$ exceeds a certain limit λ, e.g. $\lambda = 2.58$ for the approximate 99% confidence interval of the residual U.

In order to apply the reweighted structural regression (RSR), the above weights of the residuals computed by the least median structural regression of equation 10 or the absolute median structural regression of equation 11 are used, and

$$\psi_w^2(\mu_x; \mu_y; \beta)$$

$$= \sum_{i=1}^{n} w_i\{|\beta|^{-1}(y_i - \mu_y)^2 + |\beta|\,(x_i - \mu_x)^2\} \quad \text{(Eq. 12)}$$

is to be minimized with respect to the model parameters. The robust variance estimate s_v^2 of the location points v_i can be determined analogously. In accordance with Eq. 9, this makes it possible to define a robust bivariate coefficient of determination:

$$R^2 = 1 - \frac{s_u^2}{s_v^2} \quad \text{(Eq. 13)}$$

6. Application

Using a real data set we consider the comparison of two analytical methods, TOA and BGE, both measuring the packed cell volume or haematocrit, i.e. the volume of erythrocytes expressed as the fraction of the volume of whole blood in a sample. The haematocrit data are shown in the Appendix.

The aim of the comparison is to examine the accuracies of the measurement methods. Both analytical methods have the same proportional accuracy, if $\beta = 1$ holds and, furthermore, they have the same additive accuracy, if $\alpha = 0$ is valid. The test of proportional and additive bias corresponds to the question of whether or not the parameter values of the slope and the intercept are significantly different from 1 and 0.

The slope and intercept estimates and the 95%-confidence intervals as well as the coefficients of determination (Eq. 13) are shown in table 1 for the ordinary least squares regression (LS), the structural regression (SR), the absolute median structural regression (ASR), the reweighted absolute median structural regression (RSR/A) and the reweighted least median structural regression (RSR/L). For the least squares regression and absolute median structural regression the corresponding regression and residual plots are given in figures 1 and 2.

Obviously, due to several outliers, the least squares regression leads to heavily biased results (fig. 1a). Although for least squares regression *Pearson*'s product moment coefficient of correlation of the residuals, $u_i = y_i - (a + bx_i)$, and the location points, $v_i = x_i - \bar{x}$, equal zero, there is a remarkable linear trend in the residuals (fig. 1b), which is quantified by *Spearman*'s rank correlation $r_{uv} = 0.53$ in table 1.

In the structural regression models mentioned above, *Pearson*'s as well as *Spearman*'s coefficient of correlation of the residual and location points (tab. 1) do not differ significantly from zero, and hence a linear or monotone trend of the residuals with respect to their locations is not detectable in each of the bivariate methods. Hence, all bivariate methods do fit the data well. However, the estimates of the residual variance in the structural regression (Eq. 8) (see tab. 1) are obviously influenced by the outliers, and the residual variance is overestimated.

The absolute median structural regression (Eq. 11) offers an appropriate robust variance estimate of the residuals and permits the detection of outliers. In figure 2, an outlier (x_k, y_k) was defined as having a standardized robust residual greater than three, i.e. $|u_k/s_u| > 3$, and five of 112 measuring points are flagged as outliers.

In the reweighted absolute median structural regression (RSR/A) (Eq. 12) the same weights as in the absolute median structural regression (Eq. 11) were used, leading to unbiased variance estimates (tab. 1).

Additionally, the least median of squares regression (LMS) and the least median structural regression (LSR) (Eq. 10) are computed, using the simplex algorithm (16), available in the NAG-subroutine library. The least median of squares regression leads to the slope $b_x = 1.00$ and the intercept $a_x = 0.0095$ with least function value $\varphi_r^2 = 0.0001323$. The least median structural regression leads to the slope $b_x = 1.08$ and to the focal point $(m_x, m_y) = (0.331, 0.345)$ with least function value $\psi_r^2 = 0.002308$, and detects six outliers (see figs. 3a

Tab. 1. Regression and residual analysis of ordinary least squares regression (LS), structural regression (SR), reweighted absolute median structural regression (RSR/A), absolute median structural regression (ASR) and reweighted least median structural regression (RSR/L).

Type of regression		LS	SR	RSR/A	ASR	RSR/L
Regression line						
Slope	b	0.59	1.17	1.11	1.07	1.11
95% Confidence interval		(0.40, 0.77)	(0.98, 1.36)	(1.05, 1.16)	(0.94, 1.17)	(1.06, 1.17)
Intercept	a	0.163	−0.0463	−0.0233	−0.0125	−0.0250
95% Confidence interval		(0.090, 0.232)	(−0.0597, −0.0330)	(−0.0276, −0.0191)	(−0.0171, −0.0089)	(−0.0292, −0.0209)
Coefficient of determination	R^2	0.25	0.68	0.96	0.97	0.96
Residual analysis						
Standard deviation	s_u	0.0625	0.0474	0.0151	0.0139	0.0147
Standard deviation	s_v	0.0616	0.0822	0.0756	0.0752	0.0758
Correlation *Pearson*	r_{uv}	0.00	0.00	0.00	0.10	0.00
Correlation *Spearman*	r_{uv}	0.53	−0.12	−0.01	0.07	−0.03

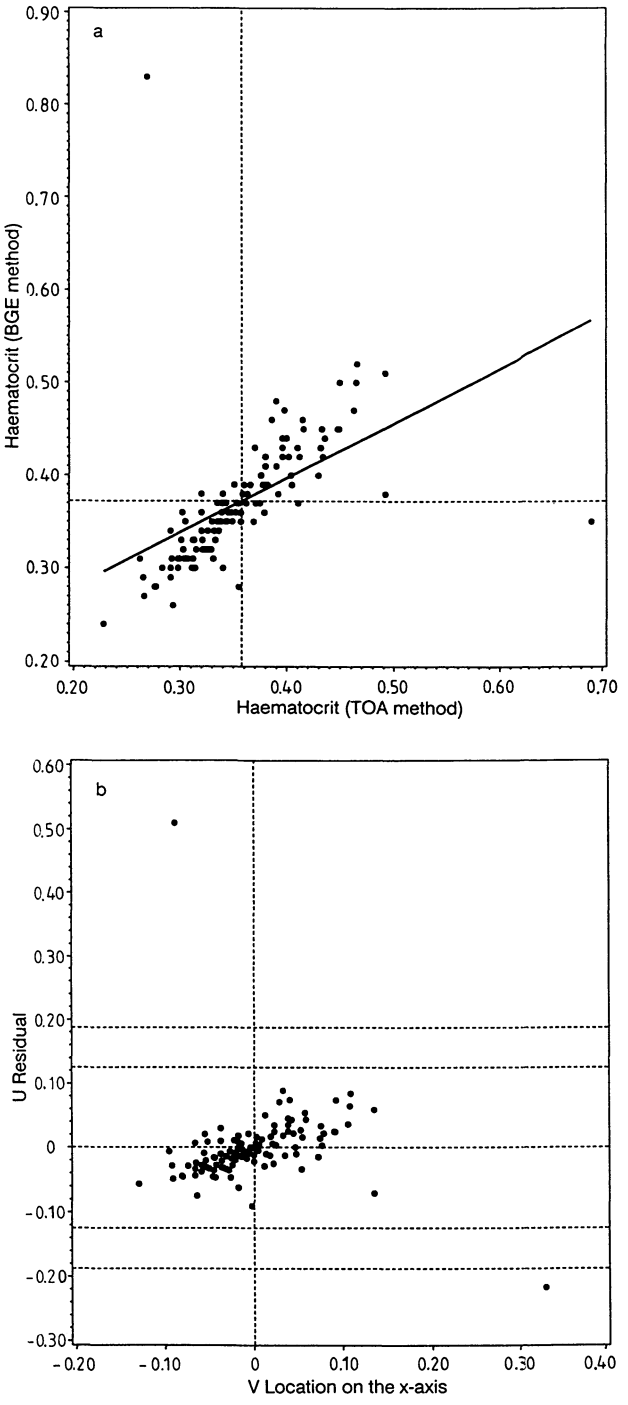
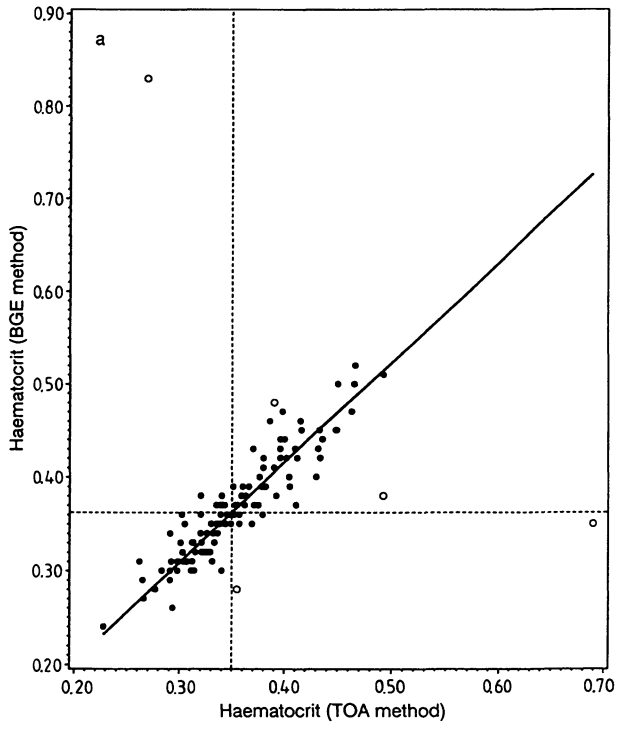


Fig. 1. Comparison of haematocrit (PCV) methods: TOA (x) vs BGE (y)
a) ordinary least squares regression. Dashed lines are the means of the measurements $PCV_{BGE} = 0.163 + 0.586 \times PCV_{TOA}$
b) residual plot. Dashed lines are $\pm 2s_u$ and $\pm 3s_u$

and 3b). The numerical results of the reweighted least median structural regression (RSR/L) are shown in table 1.

It should be noted, that the absolute median structural regression (ASR) and the least median structural regression (LSR) lead to nearly the same regression



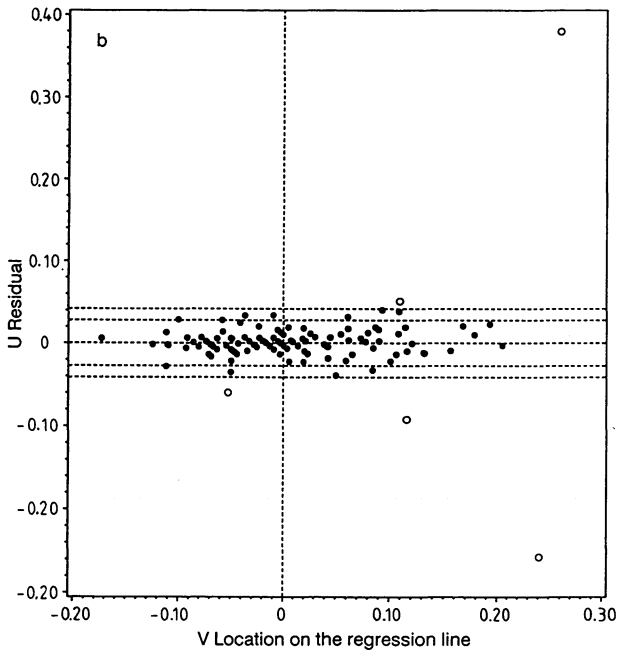
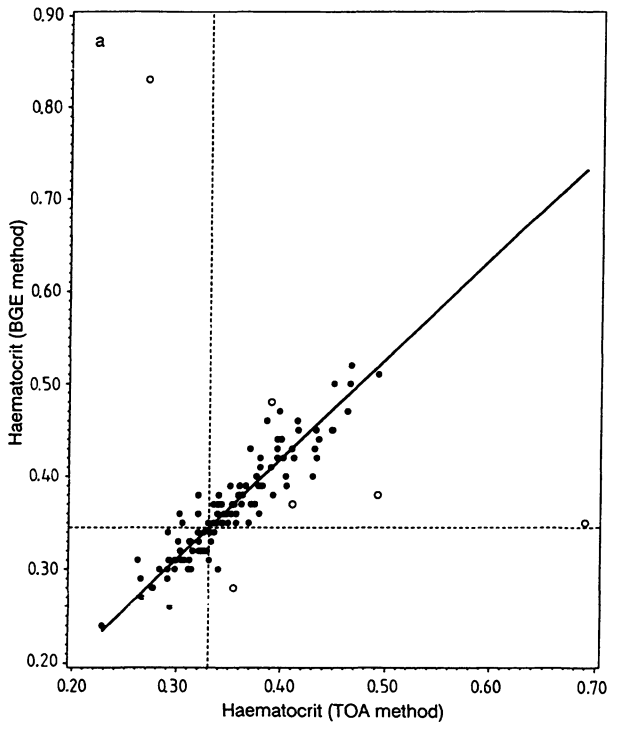


Fig. 2. Comparison of haematocrit (PCV) methods: TOA (x) vs BGE (y)
a) absolute median structural regression
Dashed lines are the robustified means of the measurements.
$PCV_{BGE} = -0.0125 + 1.07 \times PCV_{TOA}$
b) residual plot.
Dashed lines are $\pm 2s_u$ and $\pm 3s_u$.
Circles indicate outliers.



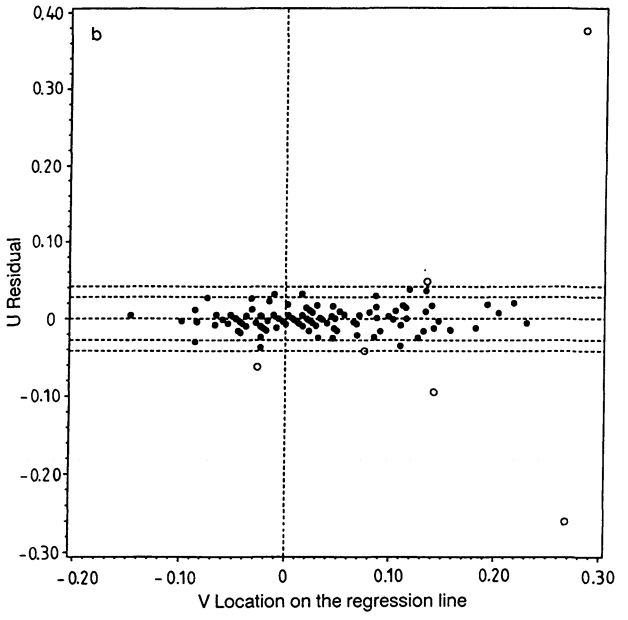


Fig. 3. Comparison of haematocrit (PCV) methods: TOA (x) vs BGE (y)
a) least median structural regression.
Dashed lines are the robustified means of the measurements.
$PCV_{BGE} = -0.012 + 1.08 \times PCV_{TOA}$
b) residual plot
Dashed lines are $\pm 2s_u$ and $\pm 3s_u$.
Circles indicate outliers.

lines (see figs. 2a and 3a), although the respective robust focal point estimates differ significantly from each other. In the least median structural regression the focal point $(m_x, m_y) = (0.331, 0.345)$ estimates the mode of the bivariate data distribution, while the absolute median structural regression yields a bivariate median focal point estimate $(m_{rx}, m_{ry}) = (0.350, 0.362)$.

If the coordinates are interchanged the results for the least median of squares regression become $b_y = 0.79$ and $a_y = 0.0613$, with $\varphi_r^2 = 0.0001152$. The least median structural regression yields $b_y = 0.92$ and $(m_y, m_x) = (0.345, 0.331)$, with $\psi_r^2 = 0.002308$. This demonstrates that the least median structural regression is equivariant against interchanging of coordinates, while the least median of squares regression produces different solutions in this situation.

Discussion

The statistical investigation of bivariate errors-in-variables has a long tradition. *Carl Friedrich Gauss* (19) published the foundations of bivariate regression and it was *Karl Pearson* (20) who developed principal component analysis. In clinical chemistry the latter technique is commonly known as *Deming*'s method. Structural relationship analysis was proposed by *Abraham Wald* (4). The extension of *Wald*'s method to robust techniques, the resistant line method, was surveyed by *Johnstone & Velleman* (8), who generalized the resistant line. *Wald* himself emphasized the essential difference between structural relationship and bivariate regression: 'The problem of finding a structural relationship must not be confused with the problem of prediction of one variable by means of the other', and he pointed out that structural relationship models may lead to biased results, if they are applied in the framework of bivariate calibration.

Structural regression analysis as outlined in this paper, seems to be an appropriate tool for the modelling of bivariate calibration. In our opinion the crucial improvement of the given bivariate regression methods is the ability to conduct the residual analysis of bivariate data, which is not available within the framework of structural relationships. In particular, tests to confirm the linearity of the bivariate regression line can be conducted in the same way as used for ordinary regression, since a residual variable as well as a location variable are both available. Finally, an appealing property of the given bivariate regression approach is the simplicity of distributional and robust parameter estimation as well as it's ability to perform outlier detection.

Acknowledgement

The valuable suggestions of Professor Dr. *B. Schneider,* which improved this paper substantially, are gratefully acknowledged. Thanks are also due to Professor Dr. *R. Haeckel* for providing haematocrit data. Special thanks are due to the Managing Editor of this journal, Professor Dr. *F. Körber*, for additional suggestions in substance and for his personal effort in publishing this paper in time for the 60th birthday of my teacher and friend *Berthold Schneider* in such an excellent form.

References

1. Huber, P. J. (1981) *Robust Statistics*. John Wiley & Sons, New York.
2. Rousseeuw, P. J. & Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.
3. Massart, D. L., Kaufmann, L., Rousseeuw, P. J. & Leroy, A. (1986) Least Median of Squares: a Robust Method for Outlier and Model Error Detection in Regression and Calibration. Anal. Chim. Acta *187*, 171–179.
4. Wald, A. (1940) The Fitting of Straight Lines if Both Variables are Subject to Error. Ann. Math. Stat. *11*, 282–300.
5. Moran, P. A. P. (1971) Estimating Structural and Functional Relationships. J. Mult. Anal. *1*, 232–255.
6. Feldmann, U., Schneider, B., Haeckel, R. & Klinkers, H. (1981) A Multivariate Approach for the Biometric Comparison of Analytical Methods in Clinical Chemistry. J. Clin. Chem. Clin. Biochem. *19*, 121–137.
7. Passing, H. & Bablok, W. (1983) A New Biometric Procedure for Testing the Equality of Measurements From Two Different Analytic Methods. J. Clin. Chem. Clin. Biochem. *21*, 709–720.
8. Johnstone, I. M. & Velleman, P. F. (1985) The Resistant Line and Related Regression Methods. J. Am. Stat. Assoc. *80*, 1041–1059.
9. Linnet, K. (1990) Estimation of the Linear Relationship Between the Measurements of Two Methods with Proportional Errors. Statistics in Medicine *9*, 1463–1473.
10. Feldmann, U. & Schneider, B. (1987) Bivariate Structural Regression Analysis: A Tool for the Comparison of Analytical Methods. Methods of Information in Medicine *26*, 205–214.
11. Freedman, D., Pisani, R. & Purves, R. (1978) *Statistics*. New York: W. W. Norton & Company.
12. Averdunk, R. & Borner, K. (1970) Korrelation der Thromboplastinzeiten bei Dicumarol-behandelten Patienten unter Verwendung verschiedener Thrombokinasepräparate. Z. Klin. Chem. Klin. Biochem. *8*, 263–268.
13. Rousseeuw, P. J. (1984) Least Median of Squares Regression. J. Am. Stat. Assoc. *79*, 871–880.
14. Steele, J. M. & Steiger, W. L. (1986) Algorithms and Complexity for Least Median of Squares. Discrete Appl. Math. *14*, 93–100.
15. Edelsbrunner, H. & Souvaine, D. L. (1990): Computing Least Median of Squares Regression Lines and Guided Topological Sweep. J. Am. Stat. Assoc. *65*, 115–119.
16. Nelder, J. A. & Mead, R. (1965) A Simple Method for Function Minimization. Computer Journal *7*, 308–313.
17. Hampel, F. R. (1975) Beyond Location Parameters: Robust Concepts and Methods. Bull. Int. Stat. Inst. *46*, 375–382.
18. Owen, D. B. (1962) *Handbook of Statistical Tables*. Addison-Wesley, Reading MA.
19. Gauss, C. F. (1821) *Theorie der den kleinsten Fehlern unterworfenen Combinationen der Beobachtungen* (Original 1821). Reprint in: Abhandlungen zur Methode der kleinsten Quadrate von C. F. Gauss. Würzburg: Physika Verlag 1964.
20. Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philos. Mag. 6th Ser. 559–572.

Prof. Dr. Uwe Feldmann
Universität Heidelberg
Klinikum Mannheim
W-6800 Mannheim 1

Appendix: Maximum likelihood estimation

Assume that sample points (x_i, y_i) for $i = 1, \ldots, n$ are independently drawn from a bivariate normal distribution (X, Y). The predicted points (x_i', y_i') are assumed to be from the same distribution and hence their density distribution function is

$$f(x_i', y_i') = \exp[-\{(x_i' - \mu_x)^2/\sigma_x^2 - 2\varrho_{xy}(x_i' - \mu_x)(y_i' - \mu_y)/\sigma_x\sigma_y) + (y_i' - \mu_y)^2/\sigma_y^2\}/\{2(1 - \varrho_{xy}^2)\}]/\gamma,$$
$$\text{with } \gamma = 2\pi\sigma_x\sigma_y\sqrt{1 - \varrho_{xy}^2}$$

The observed points (x_i, y_i) and the predicted points (x_i', y_i') are related according to Eq. 1, therefore

$$f(x_i', y_i') = \exp[-\{\beta^{-2}(y_i - \mu_y)^2/\sigma_x^2 - 2\varrho_{xy}(x_i - \mu_x)(y_i - \mu_y)/(\sigma_x\sigma_y) + \beta^2(x_i - \mu_x)^2/\sigma_y^2\}/\{2(1 - \varrho_{xy}^2)\}]/\gamma$$

According to Eq. 3b, $\sigma_x^2 = |\beta|^{-1}\sigma_x\sigma_y$ and $\sigma_y^2 = |\beta|\sigma_x\sigma_y$ hold. Note that the term $\sigma_x\sigma_y$ is independent of any particular choice of the slope β, according to Eq. 2d. This is also true for the coefficient of correlation ϱ_{xy}. Hence, equation

$$f(x_i', y_i') = \exp[-\{|\beta|^{-1}(y_i - \mu_y)^2 - 2\varrho_{xy}(x_i - \mu_x)(y_i - \mu_y) + |\beta|(X_i - \mu_x)^2\}/\{2\sigma_x\sigma_y(1 - \varrho_{xy}^2)\}]/\gamma$$

represents explicitly the dependence of the distribution function with reference to the slope β. The log-likelihood function reads

$$l(\mu_x; \mu_y; \beta) = \frac{-\{\psi^2(\mu_x; \mu_y; \beta) - 2\varrho_{xy}\psi_o(\mu_x; \mu_y)\}}{2\sigma_x\sigma_y(1 - \varrho_{xy}^2)} - n\ln(\gamma)$$

where

$$\psi^2(\mu_x; \mu_y; \beta) = \sum_{i=1}^{n} |\beta|(x_i - \mu_x)^2 + |\beta|^{-1}(y_i - \mu_y)^2$$

and

$$\psi_o(\mu_x; \mu_y) = \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

In order to get the maximum likelihood estimates of the expectations μ_x, μ_y and the slope β, the partial derivatives of the log-likelihood are calculated and set to zero:

$$\frac{\partial\, l(\mu_x; \mu_y; \beta)}{\partial\beta} = 0, \quad \frac{\partial\, l(\mu_x; \mu_y; \beta)}{\partial\mu_x} = 0, \quad \frac{\partial\, l(\mu_x; \mu_y; \beta)}{\partial\mu_y} = 0.$$

As estimates we get

$$b = \text{sign}(r_{xy})\frac{s_y}{s_x} \text{ and } m_x = \bar{x},\ m_y = \bar{y}$$

Note that these estimates are also obtained, if only the function $\psi^2(\mu_x; \mu_y; \beta)$ is minimized.

The second derivative of the log-likelihood function with respect to the slope parameter reads

$$\frac{\partial^2 l(\mu_x; \mu_y; \beta)}{\partial\beta\,\partial\beta} = \frac{-|\beta|^{-3}\sum_{i=1}^{n}(y_i - \mu_y)^2}{\sigma_x\sigma_y(1 - \varrho_{xy}^2)}$$

The mixed derivatives $\partial^2 l(\mu_x; \mu_y; \beta)/\partial\beta\,\partial\mu_x = 0$ and $\partial^2 l(\mu_x; \mu_y; \beta)/\partial\beta\,\partial\mu_y = 0$ are zero at the maximum likelihood point. According to the maximum likelihood principle, the negative inverse of the second derivative defines the variance of b after replacing the parameters by their estimates, i.e. $s_b^2 = (s_y^2/s_x^2)/(1 - r_{xy}^2)/n$.

The variance estimate of the intercept, $s_a^2 = 2s_y^2(1 - |r_{xy}|)/n$, is computed as the standard error variance of $a_i = y_i - bx_i$, by applying $s_y = |b|s_x$. Note that this is a conditional variance estimate, namely of a, given b.

Haematocrit data

x	y	x	y	x	y	x	y	x	y	x	y	x	y
0.277	0.28	0.690	0.35	0.298	0.30	0.305	0.35	0.466	0.52	0.352	0.36	0.415	0.46
0.314	0.30	0.492	0.38	0.332	0.34	0.330	0.35	0.402	0.42	0.326	0.34	0.312	0.31
0.352	0.36	0.436	0.44	0.355	0.37	0.265	0.29	0.343	0.37	0.360	0.39	0.276	0.28
0.355	0.28	0.449	0.45	0.336	0.34	0.390	0.48	0.465	0.50	0.348	0.36	0.359	0.38
0.411	0.37	0.433	0.45	0.430	0.40	0.283	0.30	0.293	0.26	0.339	0.36	0.321	0.33
0.412	0.42	0.326	0.32	0.405	0.39	0.307	0.31	0.410	0.43	0.357	0.36	0.400	0.44
0.313	0.33	0.492	0.51	0.312	0.33	0.291	0.34	0.386	0.46	0.362	0.37	0.340	0.37
0.311	0.30	0.369	0.35	0.291	0.30	0.448	0.45	0.292	0.31	0.320	0.36	0.338	0.35
0.343	0.35	0.228	0.24	0.463	0.47	0.379	0.36	0.396	0.44	0.378	0.39	0.323	0.32
0.333	0.33	0.375	0.37	0.396	0.43	0.362	0.37	0.262	0.31	0.381	0.39	0.376	0.40
0.340	0.30	0.331	0.31	0.390	0.41	0.370	0.43	0.303	0.32	0.340	0.38	0.398	0.47
0.349	0.35	0.392	0.38	0.302	0.36	0.321	0.32	0.266	0.27	0.335	0.37	0.380	0.41
0.291	0.29	0.434	0.42	0.382	0.39	0.304	0.31	0.363	0.38	0.315	0.32	0.320	0.38
0.269	0.83	0.357	0.35	0.335	0.35	0.329	0.32	0.344	0.35	0.450	0.50	0.345	0.36
0.298	0.31	0.432	0.43	0.353	0.37	0.299	0.31	0.301	0.33	0.380	0.42	0.351	0.39
0.396	0.42	0.371	0.37	0.320	0.34	0.416	0.45	0.404	0.40	0.366	0.39	0.339	0.37